DOCUMENT RESUME

ED 081 843 TM 003 204

AUTHOR Smith, Charles W.
TITLE Criterion-Referenced Assessment.
PUB DATE 17 Jul 73
NOTE 14p.; Paper presented at International Symposium cn Educational Testing (the Hague, The Netherlands, July 17, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; *Criterion Referenced Tests; Educational Testing; Measurement Techniques; *Norm Referenced Tests; Speeches; Standardized Tests; *Test Interpretation; *Test Selection
IDENTIFIERS *Mastery Learning

ABSTRACT

Both criterion-referenced and ncrm-referenced measures are useful tools to the classroom teacher, but each has its specific uses. The criterion-referenced measure is useful when one is interested in whether an individual possesses particular competencies and when there are no quotas as to how many possess that skill. It is particularly useful in assessing competence in licensed professions since tasks in these areas must be performed at specifiably high levels of competence. Criterion-referenced assessment is also important to any subject area where future academic success is dependent upon cumulative information or skills, such as in mathematics. The norm-referenced measure should be used when selectivity is required, such asin choosing the most able candidate to fill a position or when only a limited number of candidates can be selected for vocational training or academic pursuit. The criterion-referenced measure points out whether an individual possesses particular skills or competencies, but the norm-referenced measure is better able to indicate how well the individual performs in his competent area. The criterion referenced measure aims to discriminate between successive performance of a given individual, while the norm-referenced measure aims to discriminate between individuals within a particular group on a given measure. Criterion-referenced assessment, along with feedback and remedial procedures, can help teachers realize the goals of mastery learning with their students. (Author/KM)

FILMED FROM BEST AVAILABLE COPY

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 081843

TM 003 204

Criterion-Referenced Assessment

Charles W. Smith
Department of Elementary Education
Northern Illinois University
DeKalb, Illinois

A paper presented at the International Symposium on Educational Testing, the Hague, The Netherlands, July 17, 1973

Steve, in looking over his test results, found that--with his raw score of 92--he ranked at the 79th percentile rank and at the seventh stanine. His raw score had been compared with those scores obtained by his classmates. This procedure of assessing one learner's progress in relation to the performance of others in the group by using the same instrument is an example of the use of traditional norm-referenced testing. Although Steve was aware of how he ranked with his classmates, he was not provided with definitive data regarding the extent to which he met, or failed to meet, the objectives of instruction. Such testing techniques and reporting of progress are most common with standarized achievement and mental maturity testing.

Disenchantment with the measurement and reporting of pupil progress is not new among educators. A renowned psychometrist once stated:

> ...The essential fault of the older schemes for school grades or marks was that the '86' or 'B--' did not mean any objectively defined amount of knowledge or power or skill - that, for example, John's attainment of 91 in second-year German did not inform him (or anyone else) about how difficult a passage he could translate, how many words he knew the English equivalents of and how accurately he could pronounce, or about any other fact save that he was supposed to be slightly more competent than someone else marked 89...The detailed nature and the report to the individual of his school marks were not the vices of the old system. Its vice was its relativity and indefiniteness - the fact already described that a given mark did not mean any defined amount of knowledge, or power, or skill - so that it was bound to be used for

> relative achievement only...To be seventeenth instead of eighteenth, or twenty-third instead of twenty-fifth, does not approach in moving force the zeal to beat one's own record, to see one's practice curve rise week by week, and to get up to the standard which permits one to advance to a new feat.

This quotation did not come out of the 1960's but instead was written by E. L. Thorndike sixty years ago. (Thorndike, 1913)

Norm-referenced measures have been used extensively in the past in making decisions about individuals and programs based upon how students' scores compared with the scores of other students on a particular measure. The comparison may have been made regarding those who took the test locally or normative comparisions may have been utilized. Reference groups may have been determined on the basis of age level, grade level, sex, or geographical area. Such assessment measures are appropriate when selectivity among individuals is required. Though these measures point out excellence or deficiency in an individual as he is compared with others in the group, they fail to indicate what the individual can do with regard to an established standard of performance in reference to specific course objectives.

Recently instructional leaders have been giving increased attention to a variety of educational constructs. Among these are individualized instruction, continuous progress plans, non-graded programs, team teaching, humanized education, learning packets, programmed instruction, computer-assisted instruction, performance-based education, performance contracting, and accountability. Implementation of such programs and their related concepts requires that educators reevaluate the testing procedures utilized for learner and course assessment.

Criterion-referenced measures, spoken of by Glaser as early as 1963, should be seriously considered as an additional, and sometimes more appropriate, assessment measure in a variety of contemporary settings. The criterion-referenced test is designed to assess the presence or absence of criterion behaviors that have been specifically formulated from one's educational objectives. Thus information is obtained regarding the extent to which the learner has achieved the course objectives. Such information doesn't indicate how the learner compares with his peers or a normative group. These tests are appropriate where individualized instruction is stressed since they indicate whether the learner is ready to progress to the next unit of instruction. The principles of criterion-referenced testing are apparent in programmed instruction and computer-assisted instruction. In each of these situations learners proceed in a step by step manner through the instructional unit. Progression to newer concepts continues only after mastery has been indicated with previous knowledge or skills. The criterion-referenced measure can serve initially to place the learner at the appropriate instructional level. Further testing helps to diagnose accomplishments and deficiencies and gives indication of achievement.

Criterion-referenced measures serve two primary purposes. First, they provide specific information on the performance levels of individuals with regard to the instructional objectives. Second, these measures provide information that is useful in evaluating the effectiveness of instruction. The latter is more possible with the criterion-referenced measure than with the norm-referenced measure because of the closeness of item construction to the instructional objectives.

In appearance the norm-referenced measure may not differ from the criterion referenced measure. A basic difference between criterion-referenced and norm-

referenced measures is the quantitative scale used to indicate performance. In norm-referenced testing one ascertains how much the learner deviates from the average performance of the group. In criterion-referenced testing one ascertains how nearly the learner evidences a specified performance standard. Optimal scores indicate mastery of the defined abilities and scores at the bottom of the distribution indicate absence of mastery. Since test performance is described in absolute terms, the number right or the per cent correct may be an adequate means of reporting progress. With a predetermined level of performance as the goal, scoring could also be on a dichotomous scale of merely pass or fail. If percentile ranks are assigned the comparison becomes norm-referenced. Some tests might be easily enough scored as criterion-referenced and/or norm-referenced measures. The construction of such tests, however, may or may not have conformed to principles of criterion-referenced test construction.

Cartier (1968) has indicated eight additional points of contrast between the norm-referenced measure and the criterion-referenced measure. I wish to include some of my own comments along with his. Labels used for each of the eight points are my own.

(1) Variability - The norm-referenced measure is designed specifically to maximize score variability and to produce scores that are normally distributed. This is done by constructing test items primarily of medium difficulty and by striving for as wide a range of scores as possible. Both very easy items and very difficult items are minimized. The greater variability obtained leads to greater reliability. The number of order errors is thus reduced when the scores are placed in rank order. This is crucial, since norm-referenced measures are used frequently for selection purposes. Variability is not necessary or desirable in the criterion-referenced measure. A negatively skewed distribution, with a

large number of perfect or near perfect scores is expected. Variation of scores between pre- and post-test measures is desirable rather than the traditional variation between the highest and lowest scores in a particular group. Whereas norm-referenced measures attempt to maximize differences among individuals, the criterion-referenced measure is designed to discriminate between successive performances of an individual. Maximal variability typifies norm-referenced measures. Paradoxically, the effective teacher may aim to lessen variability within his/her group by facilitating learning so that all learners will demonstrate a specified level of proficiency. Such mastery is a reasonable criterion to aim for when the fundamentals of a subject area are the objectives.

(2) Scope - The norm-referenced test is likely to only sample the course objectives. The criterion-referenced test is more likely to test each essential behavior as expressed in the objectives. Several questions might be included for each objective.

(3) Style - The norm-referenced test is frequently done in an indirect manner with students answering questions about what they would do in a given situation. The criterion-referenced test is more likely to require the learner to demonstrate a behavior directly, such as repairing an engine or threading a sewing machine.

(4) Criterion - On a norm-referenced test an individual may receive a passing score by responding correctly to perhaps a third or a half of the items. On the criterion-referenced measure it is expected that the learner will answer perhaps eighty per cent or more of the items correctly. The criterion for passing is likely decided prior to test administration with the criterion-referenced test and may be decided after testing for the norm-referenced measure. Reports of progress should indicate what the learner can do and his level of proficiency.

(5) Follow-up - On formative tests of a criterion nature, there is more likely to be follow-up remedial work for each missed item, with an absolute standard or mastery as an objective.

(6) Expectations - Item writing is facilitated by stating the instructional objectives in terms of the behavioral responses that the learner is expected to be able to exhibit after instruction. The objectives and expectations for a specific test and for the course are more likely to be specifically indicated to the learner in the criterion-referenced setting. The norm-referenced test, competitive in nature because of its ranking of students, is more likely to be secretive and competitive.

(7) Missed Items - Frequent incorrect responses on a norm-referenced measure are likely to require an item revision. When an item on a criterion-referenced test is frequently missed, it is the instruction that is more likely to be questioned.

(8) Construction - Criterion-referenced measures are more difficult to construct and administer. Gronlund (1973) indicates several areas that may prove problematic to the test designer. The first of these is delimiting the tasks to be tested. As in programmed instruction, the criterion-referenced test should include items that require mastery over a restricted number of specific learning outcomes. Basic skill areas and lower level cognitive skills are most amenable to this type of test construction. A second problem is in setting performance standards. Until more empirical evidence is available, the establishment of a specified criterion level of performance remains basically a subjective judgment influenced by one's teaching experience. A normative frame of reference may be utilized. The educator may identify a likely criterion of success by studying average performance on norm-referenced measures. Crucial

to the decision is the importance of learning presently being tested to effective future learning or to on-the-job competence. Block (1971) indicated that 80 to 85 per cent mastery on formative measures is a reasonable expectation for future success in a given area. A third problem concerns sampling behavior. It's difficult to have each item representative of all possible items for a specific objective. Gronlund (1973) indicates that the classroom test will most likely do an adequate job of sampling a given area when instructional units are short, when learning tasks are specifically defined, and when procedures are employed to insure an adequate sampling.

Validity of the criterion-referenced test is judged in terms of the adequacy with which test items reflect the criterion of performance as stated in the behavioral objectives. Content validity is of prime importance. Most tests provide only a sampling of behavioral tasks. The primary consideration in the selection of any given test item for use in the criterion-referenced test is the degree to which it adequately assesses the behavior as specified in the objective. Ideally the sample of tasks required by the test will be such that one can generalize the results to that more inclusive domain of behaviors that were sampled.

Reliability, in terms of internal consistency, is an important consideration for the criterion-referenced test. All items should reflect the criterion being tested. Traditional procedures for assessing such internal consistency are inappropriate because of their dependence on score variability. Teachers might be hopeful that all their students would get perfect or near perfect scores on their criterion-referenced test. This would not contribute to score variance. Such results, if studied by traditional procedures, would yield internal consistency indices at or near zero. Popham and Husek (1969)

suggest the use of indices that reflect the ability of a test to produce variation between pre-instruction and post-instruction.

Measures of stability are equally problematic to assess, again because of the reliance of test-retest correlation coefficients upon test score variance. Livingston (1972) asserts that the farther the mean score falls from the criterion score, the greater the reliability of the test.

Lack of expected variability again interferes with the concept of item analysis as perceived traditionally. Nondiscriminating items have most frequently been thought of as those that were too easy, too difficult, or ambiguous. With instructional procedures aimed at specified levels of mastery, indices of item difficulty approach 1.00. Items on such a test, though unable to discriminate between high and low achievers, are useful when other kinds of comparisons are made such as pre-instruction vs. post-instruction. Cox and Vargas (1967) computed two discrimination indices for tests which had been administered as pre- and post-tests. One index was derived traditionally to see how well items discriminated between high and low achievers. The second index was determined by subtracting the percentage of pupils who passed the item on the pre-test from the percentage who passed the item on the post-test. The investigators concluded that some items that were found to be highly desirable on the pre-post test of discrimination were ones that would have been rejected by traditional item analysis procedures because they failed to discriminate between high and low achievers.

Although negatively discriminating items are still suspect for discarding or revision, the non-discriminating item need not be rejected. It serves a useful purpose as long as it assesses an important attribute of the criterion. While failing to discriminate between high and low achievers in the traditional

sense, such an item may still discriminate between those who have received instruction and those who have not. Difficulty level of items should not be a major concern of the writer of the criterion-referenced test. The difficulty of items should derive solely from the depth of concept being tested.

Despite some psychometric problems, several advantages of criterion-referenced testing are apparent. A primary advantage of the criterion referenced measure is found in the information it provides. Mastery of the subject is indicated to the learner, to the teacher, and to the parent in a manner that is more understandable to each. The teacher can easily evaluate the effectiveness of his instruction by analyzing the test items of his students. The teacher is also able to examine closely the learning sequence to appraise its effectivenss. This type of testing helps to assure that learners are working on learning experiences directly related to their individual goal deficiencies. Also, the type of competition fostered is with the learner himself as opposed to the pressures exerted by the competition of norm-referenced testing.

There are several reasons why criterion-referenced measures are not used more frequently. One hindrance is the time, skill, and energy required to state behavioral objectives, to choose instructional procedures that will most likely assure reaching one's objectives, and to analyze tasks to determine the types of performance that are most apt to indicate mastery or lack of it. Other construction problems have previously been mentioned. These hurdles should not be used as an excuse to avoid employing criterion-referenced measures. With dedicated effort, teachers or committees of teachers can develop effective measures for criterion-referenced testing.

The Instructional Objectives Exchange has an extensive collection of educational objectives that teachers might use as a starting point for selecting

or writing objectives. They also have criterion-referenced tests available in some areas of reading and mathematics for elementary students.

Criterion-referenced measures are most appropriate in subject areas such as math and science in which there is a hierarchy of skills. In these areas performance on one task depends upon the ability to perform previously learned tasks. It is most useful in a pre-test/post-test situation. Comparison of pre- and post-test measures gives evidence of the effectivensss of the teaching strategies utilized. These measures have also been found to be particularly effective with disadvantaged and migrant children. Since the families of these children relocate frequently, it is imperative that teachers be able to gain helpful information regarding the achievement levels of these children as soon as possible after their arrival in a new school setting.

Closely related to criterion-referenced testing is the concept of mastery learning. Advocates of mastery learning foresee success experiences for perhaps 95 per cent of their students. This follows Carroll's (1963) view that aptitude refers to the time that is necessary for a learner to gain mastery. His idea assumes that with sufficient time, proper pacing of the instructional sequence, appropriate environmental conditions, and enriched learning experiences most learners can attain mastery.

Mayo (1970) views a mastery model of learning as including (1) informing students of the course objectives prior to instruction, (2) setting a criterion of mastery prior to teaching the unit, (3) using formative tests to diagnose short-term progress, (4) prescribing additional learning experiences in those areas in which mastery is not evidenced, and (5) assuring learners sufficient time for mastery. Summative evaluation, of a criterion-referenced nature, would later indicate mastery or non-mastery at the prescribed level.

Mastery learning becomes more feasible as pace, sequence, materials and the instructional process are optimally chosen for each learner. Criterion-referenced testing fulfills a need in an era when education is more and more based on competency and during a time when schools must prove their accountability.

The criterion-referenced measure indicates what an individual can do regarding criterion behavior but it does not indicate how well he can do it with reference to others in a similar situation. In this latter situation, reliance on norm-referenced measures is again necessary.

Summary:

Both criterion-referenced and norm-referenced measures are useful tools to the classroom teacher. Each has its specific uses.

The criterion-referenced measure aims to test mastery of specified objectives in an absolute sense, not relative to any other learner's performance. It is useful when one is interested in whether an individual possesses particular competencies and when there are no quotas as to how many possess that skill. It is particulary useful in assessing competence in licensed professions since tasks in these areas must be performed at specifiably high levels of competency. Criterion-referenced assessment is also important to any subject area where future academic success is dependent upon cumulative information or skills, such as in mathematics.

The norm-referenced measure should be used when one desires to show where an individual stands with reference to other group members. This would be the case when selectivity is required in a situation, such as in choosing the most able candidate to fill a position or when only a limited number of candidates can be selected for vocational training or academic pursuit.

Whereas the criterion-referenced measure does well at pointing out whether an individual possesses particular skills or competencies, the norm-referenced measure is better able to indicate how well the individual performs in his competent area. The criterion-referenced measure aims to discriminate between successive performances of a given individual whereas the norm-referenced measure aims to discriminate between individuals within a particular group on a given measure.

Block (1971) in studying the research literature in mastery learning states that "90 per cent of the mastery learning students have achieved as well as the top 20 per cent of the non-mastery learning students." Criterion-referenced assessment, along with feedback and remedial procedures, can help you realize this goal with your students. Can you and I and other educators afford to pass it by?

REFERENCES

Block, J. H., ed. Mastery Learning: Theory and Practice. New York: Holt, Rinehart and Winston, Inc., 1971.

Bloom, B. S., ed. Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I. Cognitive Domain. New York: David McKay Company, Inc., 1956.

Bloom, B. S., et al. "Learning for Mastery," in Bloom, B. S., et al. Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book Company, 1971.

Carroll, J. B. "A Model of School Learning," Teachers College Record, 1963, 64, 723-733.

Cartier, F. A. "Criterion-Referenced Testing of Language Skills," Tesol Quarterly, 1968, 2 (1), 27-32.

Cox, R. C. and Vargas, J. S. "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests." Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois, February, 1966, ERIC 1967, Ed. 010 517.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, 1963, 18, 519-521.

Instructional Objectives Exchange, Objective Collections. Los Angeles: Instructional Objectives Exchange, 1970.

Livingston, S. A. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 1972, 9 (1), 13-26.

Mayo, S. T. "Measurement in Education: Mastery Learning and Mastery Testing," Measurement in Education, 1970, 1-4.

Popham, W. J. and Husek, T. R. "Implications of Criterion-Referenced Measurement," Journal of Educational Measurement, 1969, 6 (1), 1-9.

Popham, W. J., ed. Criterion-Referenced Measurement: An Introduction. Englewood Cliffs, New Jersey: Educational Technology Publications, Inc., 1971.

Thorndike, E. L. Educational Psychology. New York: Columbia University, 1913, 1, 286-289.